ARL-TR-7329 • JUNE 2015

US Army Research Laboratory

A Standard for Command, Control, Communications and Computers (C4) Test Data Representation to Integrate with High-Performance Data Reduction

by Ken Renard and Brian Panneton

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-7329 ● JUNE 2015

US Army Research Laboratory

A Standard for Command, Control, Communications, and Computers (C4) Test Data Representation to Integrate with High-Performance Data Reduction

by Ken Renard
Computational and Information Sciences Directorate, ARL

and

Brian Panneton
Technical and Project Engineering, LLC, Alexandria, VA

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
June 2015

2. REPORT TYPE
Final

3. DATES COVERED (From - To)
NA

4. TITLE AND SUBTITLE
A Standard for Command, Control, Communications, and Computer (C4) Test Data Representation to Integrate with High-Performance Data Reduction

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Ken Renard and Brian Panneton

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Research Laboratory
ATTN: RDRL-CIH-C
Aberdeen Proving Ground, MD 5067

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-7329

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Data reduction for analysis of Command, Control, Communications, and Computer (C4) network tests can be complex in several aspects. Test design, instrumentation configuration, instrumentation properties, failure modes, and many other factors determine how test data is interpreted and used for data reduction and further analytics. The sheer volume of data is one aspect that can be addressed with High-Performance Computing (HPC) and managed by scaling the amount of processing resources available. The US Army Research Laboratory (ARL) and Aberdeen Test Center (ATC) have developed a scalable data reduction software suite that has been successfully used for several C4 network test events. This development effort has resulted in a quick turn-around capability for reducing data suitable for analysis. During the development period, this data came from Capability Set (CS) and Instrument Calibration Events (ICE). Further operations and continued development used data sets from Network Integration Evaluation (NIE) events. This report describes one of the major challenges encountered during this development and how it was addressed so that future tests can be handled better.

15. SUBJECT TERMS
high performance computing, data reduction, network testing

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Kenneth D Renard
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER (Include area code)**
Unclassified	Unclassified	Unclassified	UU	18	410-278-4678

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

1. Introduction

Data reduction for analysis of Command, Control, Communications, and Computer (C4) network tests can be complex in several aspects. Test design, instrumentation configuration, instrumentation properties, failure modes, and many other factors determine how test data is interpreted and used for data reduction and further analytics. The sheer volume of data is one aspect that can be addressed with High-Performance Computing (HPC) and managed by scaling the amount of processing resources available. The US Army Research Laboratory (ARL) and Aberdeen Test Center (ATC) have developed a scalable data reduction software suite that has been successfully used for several C4 network test events. This development effort has resulted in a quick turn-around capability for reducing data suitable for analysis. During the development period, this data came from Capability Set (CS) and Instrument Calibration Events (ICE). Further operations and continued development used data sets from Network Integration Evaluation (NIE) events. This report describes one of the major challenges encountered during this development and how it was addressed so that future tests can be handled better.

2. Basic Components of Data Reduction Software Suite

The objective of the data reduction software suite is to populate a C4 Data Model (a set of tables) from raw test data. Raw data from ATC-instrumented systems came in the form of Binary Large Object (BLOb) files collected on Net Advanced Distributed Modular Acquisition System (NetADMAS) platforms. These files contained Global Positioning System (GPS) location, time synchronization data, and network packet data. The resulting tables focused on the correlation of packet observations at different locations on the network and low-level analysis of network and application layer protocols (Transmission Control Protocol [TCP], Voice-over Internet Protocol [VoIP], Variable Message Format [VMF], etc.). The "CommsIP" table contains a record for all packets observed on the network. Packet observations are matched as a sending and receiving observation to make a single CommsIP record, resulting in location and timing information for each datagram sent over the network. Higher-level data products such as "CommsTCPSessions" aggregate statistics to describe the performance of the TCP over the network using data from the CommsIP tables.

3. Complexities of Test Design and Instrumentation Configuration

Networks are instrumented at various points in the overall topology. These instrumentation locations are usually associated with a mobile vehicle or static tactical operating base (called a Configurable Item [CI]). Each location has an identifier and role that implies some type of local network architecture (e.g., role may be Point Of Presence [POP], Soldier Network Extension [SNE], or Tactical Communications Node [TCN]). A mapping exists between an instrumentation device identifier (ID) and the CI description. This mapping is defined by the test instrumentation team and is usually static throughout the set of events. The type of CI and role it plays has implications on how to determine the correct heuristics for interpreting raw data. This mapping requires knowledge of the test and network configuration.

Network instrumentation is usually connected via a Switched Port Analyzer (SPAN) port on a network switch. This function attempts to take copies of packets from selected interfaces and sends them out to a data collector. Due to data rates and memory/central processing unit (CPU) limitations of the switch, this process is subject to small amounts of error (mostly in the form of packets not being successfully copied to the collector device). The data collector sees the aggregate stream of copied packets without specific knowledge of what ports they were copied from or how long they may have been buffered in the switch before they were sent to the collector. In the context of a local network (such as one on a test platform) a packet has 1 of 3 “directions” associated with it:

1) **Outbound**: This packet is being sent from this local network to another local network.
2) **Inbound**: This packet is being received from another local network to a destination on the local network
3) **Local**: This packet was sent from a local network node and was destined for another local network node

When the packet is recorded on the collecting device, a heuristic must be applied to determine which direction the packet was going. This could be a simple operation, such as matching a link-layer address as the source or destination address in an Ethernet header, but it relies on knowledge of the test or network configuration. Another heuristic could be based on a known list of local Internet Protocol (IP) addresses that also relies on knowledge of the test and network configuration.

There may be multiple “tap” points in a network where performance measurements are desired. In a tactical platform, this may mean measuring the network on both “sides” of a network device such as an encryption device. This gives analysts insight into how different parts of the overall network topology are performing and where improvements are best focused. Test instrumentation devices may have multiple input ports that tag each packet observation with the port ID. These instrumentation port IDs then need to be translated to the tap point IDs before reduction and analysis can be done. This mapping is done with knowledge of the test, network, and instrumentation configuration.

As networks and test designs get more complex, there are more tap points to instrument and analyze. Unfortunately, cost considerations reduce the number of instrumentation devices available that requires aggregation of multiple tap points into a single stream of data delivered to a single port on the collection device. More complex heuristics are then required to identify multiple observations of a packet as belonging to a specific tap point. Virtual local area networks (VLANs) (802.11Q) are one method of separating packet flows that allows identification of various tap points. Given knowledge of the test and network configuration, a mapping can be made between VLAN IDs and tap points. In practice, this can become a more complex mapping due to changing VLAN IDs in various failure modes. Figure 1 shows the desired, logical layout of the network taps in a CI. Figure 2 shows a representative actual layout of the network taps where 3 different tap points all end up within the same port of the same collector. Heuristics must be applied during processing to tag each packet observation with their correct logical tap point.

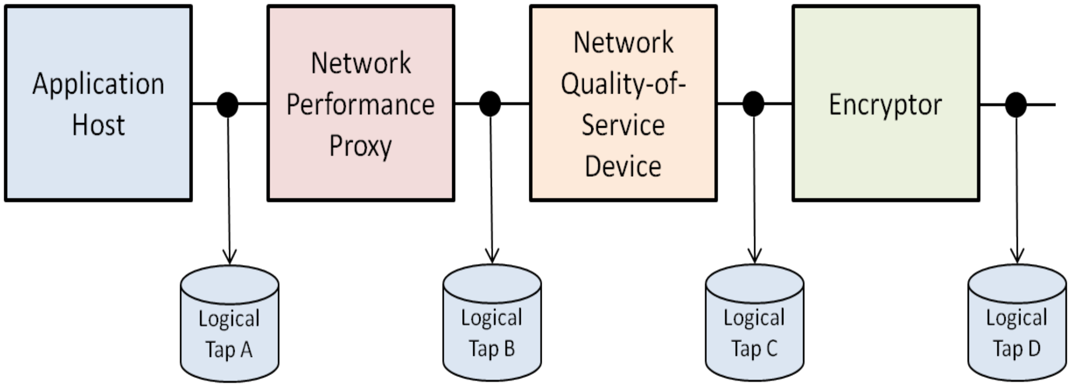

Fig. 1 Logical layout of network taps

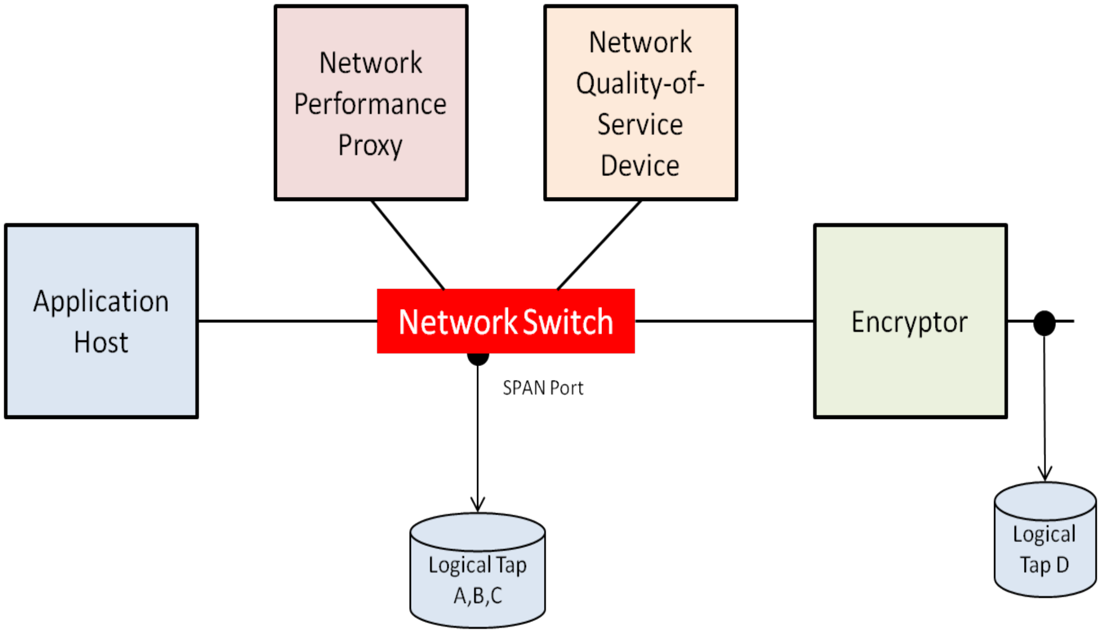

Fig. 2 Representative layout of network taps

Timing information is another type of data that must be interpreted based on selection and configuration of test instrumentation. For NetADMAS devices, time tags on packet observations are subject to variations that must be corrected by external references to time. Algorithms have been identified and implemented to combine drift-sensitive local clocks with GPS pulse-per-second sources to identify and fix clock errors. These algorithms are specific to the instrumentation platforms and must handle several types of failure modes where GPS synchronization is lost, timing records span BLOb file boundaries, or time records have bit errors.

The set of test-, network-, and instrumentation-specific information and heuristic selection necessary to properly interpret data for reduction is collectively known as "context" information. Context information is derived from test plans, network diagrams, and intimate knowledge of the test architectures and instrumentation. During the design, setup, and execution of the test, any of the information contained in the context can change, sometimes without notification, based on a variety of factors (failed instrumentation, vehicle reassignment, accidental changes in the field, etc.). Human error in recording or maintaining context data sources introduce error. This information that is critical to the interpretation of test data is therefore highly prone to error.

When processing raw data, it is important to have accurate context data so that packet observations are correctly identified and the proper correlations are made. Selection of heuristics for packet direction, tap identification, time correction, and device identification depend on this input and significant errors in data products can result from incorrect or missing context information.

4. Previous Approach

The initial implementation of the data reduction software suite implemented heuristics in the core of the code to preserve a steady data flow of raw data into the high-speed processing engine. Device identification, packet direction, and tap points were determined as packets were read in from BLOb files for processing (see Fig. 3). While this lead to an efficient load distribution among available processors, context errors or omissions were not able to be identified until incorrect data products were generated. Only upon inspection of results were subtle or obvious clues visible to indicate errors and that the processing had to be re-executed with a corrected context.

Often, the heuristics identified during design and pretest trials needed to be changed to reflect the reality of the actual tests. As testing architecture and instrumentation changes took place, assumptions made based on intended test design were proven to not hold for actual test runs. This led to last-minute changes to the core of the reduction code to replace, enhance, or change heuristics. The maintenance of the core code throughout the test events was ad-hoc and problematic due to time constraints and changing requirements.

Determining errors in context and heuristics required expertise from the HPC and testing teams to dig deep into the core of the code to identify, fix, and validate changes. This resource drain became problematic during peak activity periods of a test event.

A redistribution of functionality within the overall reduction framework was determined to be the best course of action for future generations of HPC code. The complexities that involve test-specific knowledge would be moved out of the core of the HPC processing code into pre and postprocessing phases. This would allow for identification of some set of context errors before HPC processing takes place, and for validation of context-driven decisions outside the scope of an HPC job. Given a set of inputs with context decisions and test-specific parameters already applied, the HPC codes can be more stable and focus on performance.

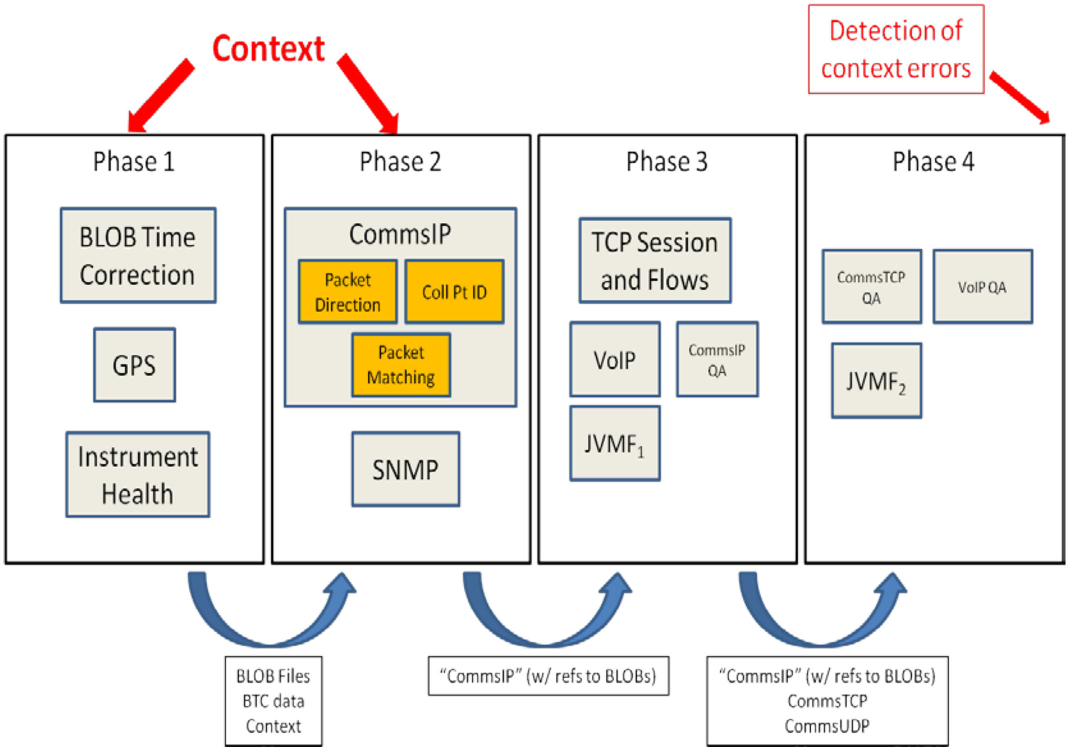

Fig. 3 Previous use of context data in HPC processing phases

5. Current Approach

The current approach relies on data being presented to the data reduction software where it is already tagged with the device ID, direction, and tap point and any necessary time correction has already been applied. This allows for a separation of the context application and the core reduction functions so that each can be validated and constructed independently. The test familiarity and expertise required for context functions is then separate from the HPC processing and its complexity so that appropriate personnel can make focused efforts in their areas of expertise independently of each other (see Fig. 4).

The standard format for presenting data to this data reduction framework is now PCAP files (see Fig. 5). The BLOb file, with its metadata, is time-corrected, and split into separate files based on Device, Tap, and Direction (DTD). The resulting PCAP files can cover an arbitrary time span, but packets must be time-ordered within the file, and separate PCAP files with the same DTD may not contain overlapping time periods of packets. Specifying the DTD for a particular PCAP file can be done with a filename convention or with a simple table of filename-to-DTD mapping (e.g., Comma-Separated Value [CSV], JavaScript Object Notation [JSON], or Structured Query Language, Lite [SQLite]).

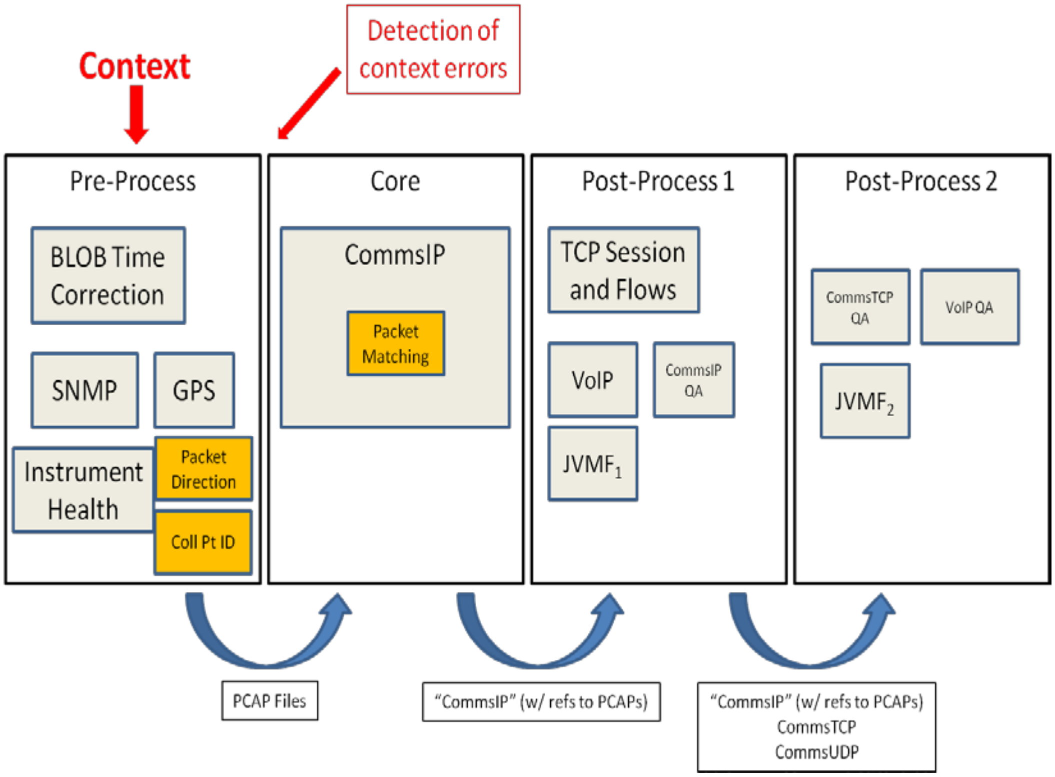

Fig. 4 Current use of context data in HPC processing phases

Heuristics for DTD determination are removed from the core HPC code (in the CommsIP module) and can be used as examples for the new DTD preprocessing phase. It is expected that verification and validation tools can be built and used on the preprocessed data to assure that the heuristics are being applied correctly, and that the results match the desired interpretation of the test. These tools would be used before an HPC reduction job is run, or even in parallel such that the reduction could start and only be stopped if a context error is found.

6. Future Standard Formats

As DTD intelligence and tagging is moved closer to the field devices, packet data in single files could be tagged individually as they are collected. The "PcapNg"[1] file format would be a good fit to support this type of tagging, while also enabling higher-precision time stamping. As of late 2014, implementations of this file format were not available to support the type of tagging needed. If there is enough interest in using the format for testing/reduction formats as well as other users in the Department of Defense community, it would be worthwhile to contribute to the implementations and standards development.

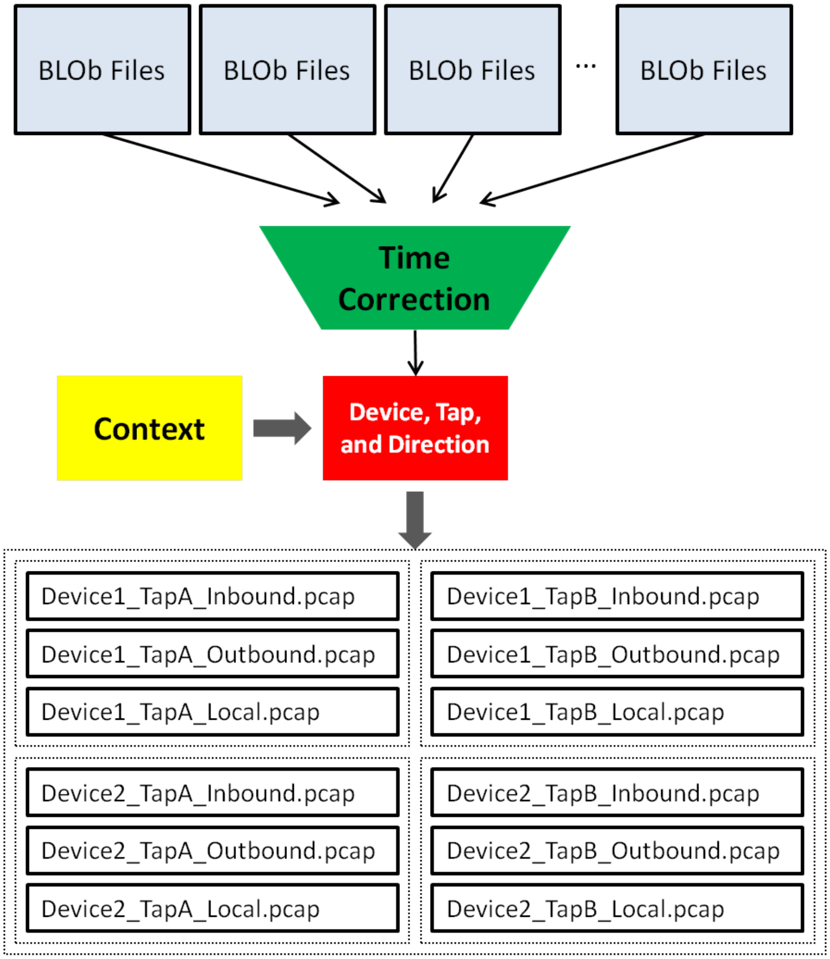
BLOb Files
BLOb Files
BLOb Files
…
BLOb Files
Time Correction
Context
Device, Tap, and Direction
Device1_TapA_Inbound.pcap
Device1_TapA_Outbound.pcap
Device1_TapA_Local.pcap
Device1_TapB_Inbound.pcap
Device1_TapB_Outbound.pcap
Device1_TapB_Local.pcap
Device2_TapA_Inbound.pcap
Device2_TapA_Outbound.pcap
Device2_TapA_Local.pcap
Device2_TapB_Inbound.pcap
Device2_TapB_Outbound.pcap
Device2_TapB_Local.pcap

Fig. 5 Standard for presentation of packet data to data reduction code

7. Notes

1. Additional information can be found at:

   https://wiki.wireshark.org/Development/PcapNg

List of Symbols, Abbreviation, and Acronyms

ARL	US Army Research Laboratory
ATC	Aberdeen Test Center
BLOb	Binary Large Object
C4	Command, Control, Communications, and Computer
CI	Configurable Item
CPU	central processing unit
CS	Capability Set
CSV	Comma-Separated Value
DTD	Device, Tap, and Direction
GPS	Global Positioning System
HPC	High-Performance Computing
ICE	Instrument Calibration Events
ID	identifier
IP	Internet Protocol
JSON	JavaScript Object Notation
NetADMAS	Net Advanced Distributed Modular Acquisition System
NIE	Network Integration Evaluation
POP	Point of Presence
SNE	Soldier Network Extension
SPAN	Switched Port Analyzer
SQLite	Structured Query Language, Lite
TCN	Tactical Communications Node
TCP	Transmission Control Protocol
VLAN	virtual local area network
VMF	Variable Message Format
VoIP	Voice-over Internet Protocol

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

2 DIRECTOR
(PDF) US ARMY RESEARCH LAB
RDRL CIO LL
IMAL HRA MAIL & RECORDS
MGMT

1 GOVT PRINTG OFC
(PDF) A MALHOTRA

1 DIRECTOR
(PDF) US ARMY RESEARCH LAB
RDRL CIH C
K D RENARD
B PANNETON

1 ATC
(PDF) J R FEIGHT

INTENTIONALLY LEFT BLANK.